Programming Fundamentals for Data Science

Laboratory Session 3

Data Processing

This laboratory session is focused on a range of data processing techniques. We will be using the Python libraries NumPy and pandas, which provide key data structures, such as arrays, series and data frames.

Task 1

Initialise an array with 12 random integer values from the interval [65 .. 75]. Visualise the values. Consider these values as the average temperatures measured in Fahrenheit. Store the values in a pandas Series. Set the indices of the pandas series as the full names of the months. Visualise the series including the indices. Convert the temperatures from Fahrenheit into Celsius and visualise them again. Use integer data type. Mode is the value which occurs most frequently in a dataset. By using the function with the same name, determine the most common average monthly temperature and print it on the screen. Print the number of months this average temperature was recorded.

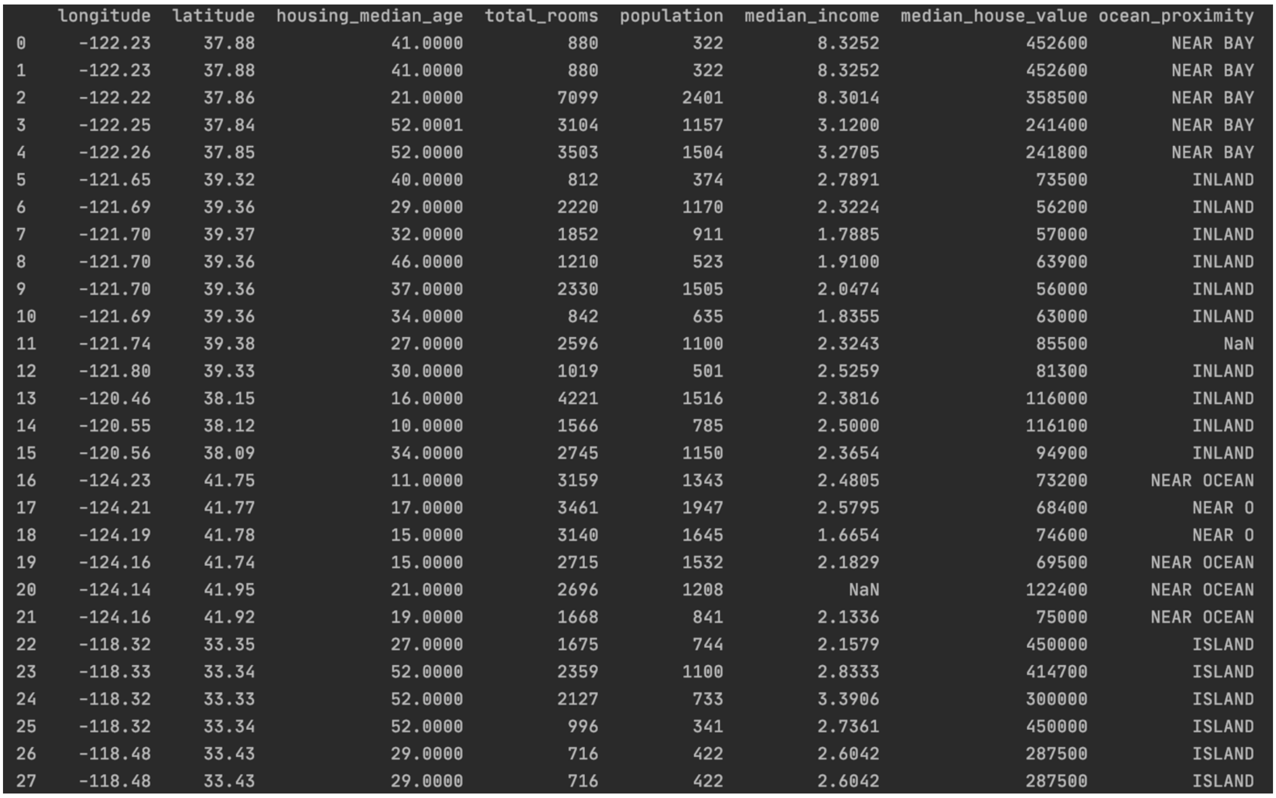
The script will need the following libraries:

```python
import numpy as np
import pandas as pd
import calendar
```

Data Frames

In Python, the term data frame describes a two-dimensional data structure organised and presented as a table. It is provided by the pandas library. In many cases in practice, the initial step to perform when working with data is the so-called data pre-processing, which is usually focused on missing values, unnecessary duplicates, wrong data types and incorrect values.

Consider the following 28-row dataset stored as 'Dataset.csv' file available on Moodle:



```
    longitude  latitude  housing_median_age  total_rooms  population  median_income  median_house_value ocean_proximity
0     -122.23     37.88             41.0000          880         322         8.3252              452600        NEAR BAY
1     -122.23     37.88             41.0000          880         322         8.3252              452600        NEAR BAY
2     -122.22     37.86             21.0000         7099        2401         8.3014              358500        NEAR BAY
3     -122.25     37.84             52.0001         3104        1157         3.1200              241400        NEAR BAY
4     -122.26     37.85             52.0000         3503        1504         3.2705              241800        NEAR BAY
5     -121.65     39.32             40.0000          812         374         2.7891               73500          INLAND
6     -121.69     39.36             29.0000         2220        1170         2.3224               56200          INLAND
7     -121.70     39.37             32.0000         1852         911         1.7885               57000          INLAND
8     -121.70     39.36             46.0000         1210         523         1.9100               63900          INLAND
9     -121.70     39.36             37.0000         2330        1505         2.0474               56000          INLAND
10    -121.69     39.36             34.0000          842         635         1.8355               63000          INLAND
11    -121.74     39.38             27.0000         2596        1100         2.3243               85500             NaN
12    -121.80     39.33             30.0000         1019         501         2.5259               81300          INLAND
13    -120.46     38.15             16.0000         4221        1516         2.3816              116000          INLAND
14    -120.55     38.12             10.0000         1566         785         2.5000              116100          INLAND
15    -120.56     38.09             34.0000         2745        1150         2.3654               94900          INLAND
16    -124.23     41.75             11.0000         3159        1343         2.4805               73200      NEAR OCEAN
17    -124.21     41.77             17.0000         3461        1947         2.5795               68400          NEAR O
18    -124.19     41.78             15.0000         3140        1645         1.6654               74600          NEAR O
19    -124.16     41.74             15.0000         2715        1532         2.1829               69500      NEAR OCEAN
20    -124.14     41.95             21.0000         2696        1208            NaN              122400      NEAR OCEAN
21    -124.16     41.92             19.0000         1668         841         2.1336               75000      NEAR OCEAN
22    -118.32     33.35             27.0000         1675         744         2.1579              450000          ISLAND
23    -118.33     33.34             52.0000         2359        1100         2.8333              414700          ISLAND
24    -118.32     33.33             52.0000         2127         733         3.3906              300000          ISLAND
25    -118.32     33.34             52.0000          996         341         2.7361              450000          ISLAND
26    -118.48     33.43             29.0000          716         422         2.6042              287500          ISLAND
27    -118.48     33.43             29.0000          716         422         2.6042              287500          ISLAND
```

Actual dataset has approximately 20,000 rows and can be found at:
https://www.kaggle.com/camnugent/california-housing-prices

Task 2

1. Download the Dataset.csv file from Moodle, load its content into pandas data frame and visualise the entire content of the data frame

Check the data frame for the following 'data cleaning' issues and resolve them:

2. Missing values
3. Unnecessary duplicates
4. Wrong data types
5. Wrong values
6. Save the updated data frame into a new CSV file

The house prices are at the focus of this data frame. By using the updated data frame, provide the following values, which describe the column 'median_house_value':

7. Mean
8. Median
9. Range
10. The column 'median_income' contains currency in tens of thousands USD. Convert it into USD and visualise the entire updated data frame